METHOD AND APPARATUS FOR IMPROVING THE INTELLIGIBILITY OF DIGITALLY COMPRESSED SPEECH

TECHNICAL FIELD

The invention relates generally to speech processing and, more particularly, to techniques for enhancing the intelligibility of processed speech.

BACKGROUND OF THE INVENTION

Human speech generally has a relatively large dynamic range. For example, the amplitudes of some consonant sounds (e.g., the unvoiced consonants P, T, S, and F) are often 30 dB lower than the amplitudes of vowel sounds in the same spoken sentence. Therefore, the consonant sounds will sometimes drop below a listener's speech detection threshold, thus compromising the intelligibility of the speech. This problem is exacerbated when the listener is hard of hearing, the listener is located in a noisy environment, or the listener is located in an area that receives a low signal strength.

Traditionally, the potential unintelligibility of certain sounds in a speech signal was overcome using some form of amplitude compression on the signal. For example, in one prior approach, the amplitude peaks of a speech signal were clipped and the resulting signal was amplified so that the difference between the peaks of the new signal and the low

portions of the new signal would be reduced while maintaining the signal's original loudness. Amplitude compression, however, often leads to other forms of distortion within the resultant signal, such as the harmonic distortion resulting from flattening out the high amplitude components of the signal. In addition, amplitude compression techniques tend to amplify some undesired low-level signal components (e.g., background noise) in an inappropriate manner, thus compromising the quality of the resultant signal.

Therefore, there is a need for a method and apparatus that is capable of enhancing the intelligibility of processed speech without the undesirable effects associated with prior techniques.

SUMMARY OF THE INVENTION

The present invention relates to a system that is capable of significantly enhancing the intelligibility of processed speech. The system first divides the speech signal into frames or segments as is commonly performed in certain low bit rate speech encoding algorithms, such as Linear Predictive Coding (LPC) and Code Excited Linear Prediction (CELP). The system then analyzes the spectral content of each frame to determine a sound type associated with that frame. The analysis of each frame will typically be performed in the

context of one or more other frames surrounding the frame of interest. The analysis may determine, for example, whether the sound associated with the frame is a vowel sound, a voiced fricative, or an unvoiced plosive.

Based on the sound type associated with a particular frame, the system will then modify the frame if it is believed that such modification will enhance intelligibility. For example, it is known that unvoiced plosive sounds commonly have lower amplitudes than other sounds within human speech. The amplitudes of frames identified as including unvoiced plosives are therefore boosted with respect to other frames. In addition to modifying a frame based on the sound type associated with that frame, the system may also modify frames surrounding that particular frame based on the sound type associated with the frame. For example, if a frame of interest is identified as including an unvoiced plosive, the amplitude of the frame preceding this frame of interest can be reduced to ensure that the plosive isn't mistaken for a spectrally similar fricative. By basing frame modification decisions on the type of speech included within a particular frame, the problems created by blind signal modifications based on amplitude (e.g., boosting all low-level signals) are avoided. That is, the inventive principles allow frames to be

modified selectively and intelligently to achieve an enhanced signal intelligibility.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram illustrating a speech processing system in accordance with one embodiment of the present invention;

Fig. 2 is a flowchart illustrating a method for processing a speech signal in accordance with one embodiment of the invention; and

Figs. 3 and 4 are portions of a flowchart illustrating a method for use in enhancing the intelligibility of speech signals in accordance with one embodiment of the present invention.

DETAILED DESCRIPTION

The present invention relates to a system that is capable of significantly enhancing the intelligibility of processed speech. The system determines a sound type associated with individual frames of a speech signal and modifies those frames based on the corresponding sound type. In one approach, the inventive principles are implemented as an enhancement to well-known speech encoding algorithms, such as the LPC and CELP algorithms, that perform frame-based speech digitization.

The system is capable of improving the intelligibility of speech signals without generating the distortions often associated with prior art amplitude clipping techniques. The inventive principles can be used in a variety of speech applications including, for example, messaging systems, IVR applications, and wireless telephone systems. The inventive principles can also be implemented in devices designed to aid the hard of hearing such as, for example, hearing aids and cochlear implants.

Fig. 1 is a block diagram illustrating a speech processing system 10 in accordance with one embodiment of the present invention. The speech processing system 10 receives an analog speech signal at an input port 12 and converts this signal to a compressed digital speech signal which is output at an output port 14. In addition to performing signal compression and analog to digital conversion functions on the input signal, the system 10 also enhances the intelligibility of the input signal for later playback. As illustrated, the speech processing system 10 includes: an analog to digital (A/D) converter 16, a frame separation unit 18, a frame analysis unit 20, a frame modification unit 22, and a compression unit 24. It should be appreciated that the blocks illustrated in Fig. 1 are functional in nature and do not

necessarily correspond to discrete hardware elements. In one embodiment, for example, the speech processing system 10 is implemented within a single digital processing device. Hardware implementations, however, are also possible.

With reference to Fig. 1, the analog speech signal received at port 12 is first sampled and digitized within the A/D converter 16 to generate a digital waveform for delivery to the frame separation unit 18. The frame separation unit 18 is operative for dividing the digital waveform into individual time-based frames. In a preferred approach, these frames are each about 20 to 25 milliseconds in length. The frame analysis unit 20 receives the frames from the frame separation unit 18 and performs a spectral analysis on each individual frame to determine a spectral content of the frame. The frame analysis unit 20 then transfers each frame's spectral information to the frame modification unit 22. The frame modification unit 22 uses the results of the spectral analysis to determine a sound type (or type of speech) associated with each individual frame. The frame modification unit 22 then modifies selected frames based on the identified sound types. The frame modification unit 22 will normally analyze the spectral information corresponding to a frame of interest and also the spectral information corresponding to one or more

frames surrounding the frame of interest to determine a sound type associated with the frame of interest.

The frame modification unit 22 includes a set of rules for modifying selected frames based on the sound type associated therewith. In one embodiment, the frame modification unit 22 also includes rules for modifying frames surrounding a frame of interest based on the sound type associated with the frame of interest. The rules used by the frame modification unit 22 are designed to increase the intelligibility of the output signal generated by the system 10. Thus, the modifications are intended to emphasize the characteristics of particular sounds that allow those sounds to be distinguished from other similar sounds by the human ear. Many of the frames may remain unmodified by the frame modification unit 22 depending upon the specific rules programmed therein.

The modified and unmodified frame information is next transferred to the data assembly unit 24 which assembles the spectral information for all of the frames to generate the compressed output signal at output port 14. The compressed output signal can then be transferred to a remote location via a communication medium or stored for later decoding and playback. It should be appreciated that the intelligibility

enhancement functions of the frame modification unit 22 of Fig. 1 can alternatively (or additionally) be performed as part of the decoding process during signal playback.

In one embodiment, the inventive principles are implemented as an enhancement to certain well-known speech encoding and/or decoding algorithms, such as the Linear Predictive Coding (LPC) algorithm and the Code-Excited Linear Prediction (CELP) algorithm. In fact, the inventive principles can be used in conjunction with virtually any encoding or decoding algorithm that is based upon frame-based speech digitization (i.e., breaking up speech into individual time-based frames and then capturing the spectral content of each frame to generate a digital representation of the speech). Typically, these algorithms utilize a mathematical model of human vocal tract physiology to describe each frame's spectral content in terms of human speech mechanism analogs, such as overall amplitude, whether the frame's sound is voiced or unvoiced, and, if the sound is voiced, the pitch of the sound. This spectral information is then assembled into a compressed digital speech signal. A more detailed description of various speech digitization algorithms that can be modified in accordance with the present invention can be found in the paper "Speech Digitization and Compression" by Paul Michaelis,

International Encyclopedia of Ergonomics and Human Factors, edited by Waldamar Karwowski, published by Taylor & Francis, London, 2000, which is hereby incorporated by reference.

In accordance with one embodiment of the invention, the spectral information generated within such algorithms (and possibly other spectral information) is used to determine a sound type associated with each frame. Knowledge about which sound types are important for intelligibility and are typically harder to hear is then used to develop rules for modifying the frame information in a manner that increases intelligibility. The rules are then used to modify the frame information of selected frames based on the determined sound type. The spectral information for each of the frames, whether modified or unmodified, is then used to develop the compressed speech signal in a conventional manner (e.g., the manner typically used by the LPC, CELP, or other similar algorithms).

Fig. 2 is a flowchart illustrating a method for processing an analog speech signal in accordance with one embodiment of the present invention. First, the speech signal is digitized and separated into individual frames (step 30). A spectral analysis is then performed on each individual frame to determine a spectral content of the frame (step 32).

Typically, spectral parameters such as amplitude, voicing, and pitch (if any) of sounds will be measured during the spectral analysis. The spectral content of the frames is next analyzed to determine a sound type associated with each frame (step 34). To determine the sound type associated with a particular frame, the spectral content of other frames surrounding the particular frame will often be considered. Based on the sound type associated with a frame, information corresponding to the frame may be modified to improve the intelligibility of the output signal (step 36). Information corresponding to frames surrounding a frame of interest may also be modified based on the sound type of the frame of interest. Typically, the modification of the frame information will include boosting or reducing the amplitude of the corresponding frame. However, other modification techniques are also possible. For example, the reflection coefficients that govern spectral filtering can be modified in accordance with the present invention. The spectral information corresponding to the frames, whether modified or unmodified, is then assembled into a compressed speech signal (step 38). This compressed speech signal can later be decoded to generate an audible speech signal having enhanced intelligibility.

Figs. 3 and 4 are portions of a flowchart illustrating a method for use in enhancing the intelligibility of speech signals in accordance with one embodiment of the present invention. The method is operative for identifying unvoiced fricatives and voiced and unvoiced plosives within a speech signal and for adjusting the amplitudes of corresponding frames of the speech signal to enhance intelligibility. Unvoiced fricatives and unvoiced plosives are sounds that are typically lower in volume in a speech signal than other sounds in the signal. In addition, these sounds are usually very important to the intelligibility of the underlying speech. A voiced speech sound is one that is produced by tensing the vocal cords while exhaling, thus giving the sound a specific pitch caused by vocal cord vibration. The spectrum of a voiced speech sound therefore includes a fundamental pitch and harmonics thereof. An unvoiced speech sound is one that is produced by audible turbulence in the vocal tract and for which the vocal cords remain relaxed. The spectrum of an unvoiced speech signal is typically similar to that of white noise.

With reference to Fig. 3, an analog speech signal is first received (step 50) and then digitized (step 52). The digital waveform is then separated into individual frames

(step 54). In a preferred approach, these frames are each about 20 to 25 milliseconds in length. A frame-by-frame analysis is then performed to extract and encode data from the frames, such as amplitude, voicing, pitch, and spectral filtering data (step 56). When the extracted data indicates that a frame includes an unvoiced fricative, the amplitude of that frame is increased in a manner that is designed to increase the likelihood that the loudness of the sound in a resulting speech signal exceeds a listener's detection threshold (step 58). The amplitude of the frame can be increased, for example, by a predetermined gain value, to a predetermined amplitude value, or the amplitude can be increased by an amount that depends upon the amplitudes of the other frames within the same speech signal. A fricative sound is produced by forcing air from the lungs through a constriction in the vocal tract that generates audible turbulence. Examples of unvoiced fricatives include the "f" in fat, the "s" in sat, and the "ch" in chat. Fricative sounds are characterized by a relatively constant amplitude over multiple sample periods. Thus, an unvoiced fricative can be identified by comparing the amplitudes of multiple successive frames after a decision has been made that the frames correspond to unvoiced sounds.

When the extracted data indicates that a frame is the initial component of a voiced plosive, the amplitude of the frame preceding the voiced plosive is reduced (step 60). A plosive is a sound that is produced by the complete stoppage and then sudden release of the breath. Plosive sounds are thus characterized by a sudden drop in amplitude followed by a sudden rise in amplitude within a speech signal. An example of voiced plosives includes the "b" in bait, the "d" in date, and the "g" in gate. Plosives are identified within a speech signal by comparing the amplitudes of adjacent frames in the signal. By decreasing the amplitude of the frame preceding the voiced plosive, the amplitude "spike" that characterizes plosive sounds is accentuated, resulting in enhanced intelligibility.

When the extracted data indicates that a frame is the initial component of an unvoiced plosive, the amplitude of the frame preceding the unvoiced plosive is decreased and the amplitude on the frame including the unvoiced plosive is increased (step 62). The amplitude of the frame preceding the unvoiced plosive is decreased to emphasize the amplitude "spike" of the plosive as described above. The amplitude of the frame including the initial component of the unvoiced plosive is increased to increase the likelihood that the

loudness of the sound in a resulting speech signal exceeds a listener's detection threshold.

With reference to Fig. 4, a frame-by-frame reconstruction of the digital waveform is next performed using, for example, the amplitude, voicing, pitch, and spectral filtering data (step 64). The individual frames are then concatenated into a complete digital sequence (step 66). A digital to analog conversion is then performed to generate an analog output signal (step 68). The method illustrated in Figs. 4 and 5 can be performed all at one time as part of a real-time intelligibility enhancement procedure or it can be performed in multiple sub-procedures at different times. For example, if the method is implemented within a hearing aid, the entire method will be used to transform an input analog speech signal into an enhanced output analog speech signal for detection by a user of the hearing aid. In an alternative implementation, steps 50 through 62 may be performed as part of a speech signal encoding procedure while steps 64 through 68 are performed as part of a subsequent speech signal decoding procedure. In another alternative implementation, steps 50 through 56 are performed as part of a speech signal encoding procedure while steps 58 through 68 are performed as part of a subsequent speech decoding procedure. In the period between

the encoding procedure and the decoding procedure, the speech signal can be stored within a memory unit or be transferred between remote locations via a communication channel. In a preferred implementation, steps 50 through 56 are performed using well-known LPC or CELP encoding techniques. Similarly, steps 64 through 68 are preferably performed using well-known LPC or CELP decoding techniques.

In a similar manner to that described above, the inventive principles can be used to enhance the intelligibility of other sound types. Once it has been determined that a particular type of sound presents an intelligibility problem, it is next determined how that type of sound can be identified within a frame of a speech signal (e.g., through the use of spectral analysis techniques and comparisons between adjacent frames). It is then determined how a frame including such a sound needs to be modified to enhance the intelligibility of the sound when the compressed signal is later decoded and played back. Typically, the modification will include a simple boosting of the amplitude of the corresponding frame, although other types of frame modification are also possible in accordance with the present invention (e.g., modifications to the reflection coefficients that govern spectral filtering).

An important feature of the present invention is that compressed speech signals generated using the inventive principles can usually be decoded using conventional decoders (e.g., LPC of CELP decoders) that have not been modified in accordance with the invention. In addition, decoders that have been modified in accordance with the present invention can also be used to decode compressed speech signals that were generated without using the principles of the present invention. Thus, systems using the inventive techniques can be upgraded piecemeal in an economical fashion without concern about widespread signal incompatibility within the system.

Although the present invention has been described in conjunction with its preferred embodiments, it is to be understood that modifications and variations may be resorted to without departing from the spirit and scope of the invention as those skilled in the art readily understand. Such modifications and variations are considered to be within the purview and scope of the invention and the appended claims.